(* TWO-WAY *)

23

Fitting Can Be Confusing

(posses and descriptions)

[essential background]

		Exh.	Page
Introduction	23-1	1 a	23-4
		2	23-5
23A.	A basic remark 23-1	3	23-7/8
23B.	A variable can appear repeatedly 23-6	4 b	23-12
	- two carriers can be near surrogates 23-10 to 23-13	5	23-14
	- a constant can be a surrogate 23-13 to 23-15	c	
		6 d	23-23
23C.	Posses 23-15	7	23-25
23D.	Writing the same thing down two ways 23-17	e	
	- what do the details mean? 23-19 to 23-21	f	
	- "held fixed", you say? 23-21 to 23-23	g	
	- the third caution 23-24	p	
	- the three shadows 23-24 to 23-27		
23E.	But is this exploratory? 23-27		
23F.	Unplanned data 23-29		
	* * *		
23H.	How far have we come? 23-32		
	* * *		

23

Fitting Can Be Confusing

(posses and descriptions)

[essential background]

In the last chapter, we came to realize that formulas can help us understand what we are fitting. We saw two sorts of examples:

- cases where formulas and simple algebra made clear the equivalence of things which, if described in words, seem far from obviously equivalent.
- cases where a sequence of alternative fits can be much more clearly related to one another if we use formulas.

We also saw examples where, though we avoided pressing the point, the data is weaker than we might think in telling us about certain aspects of certain fits.

We want to push these matters further in this chapter, and -- as well -- we want to relate other simple fitting problems to two-way tables. (Only quite simple problems will do, since even slightly complex fitting questions turn out to correspond to three- and more-way tables.)

23A. A basic remark

Suppose that we have data involving:

- a response, y,
- one factor, x,
- a second factor, t,

and that we are prepared to fit

straight-line dependences

on any one of three choices of carriers:

- x alone,
- t alone,
- x and t together,

and that:

(*) the relationship of x to t -- for the data we have -- is both close and nearly straight-line.

This situation arises all too frequently. We need to understand the nature and strength of any difficulties that the nearly linear relation between x and t will cause us. Our greatest concern is with cases where:

- taking a tilt on x out of y is quite effective -- in the sense that the residuals this leaves are markedly more alike than were the values of y.

In view of (*), each such case is also one where:

- taking a tilt on t out of y is quite effective -- in the sense that the residuals this leaves are markedly more alike than were the values of y.

There is no escape. If fitting a straight-line dependence on x helps a lot, so does fitting a straight-line dependence on t . This is an inevitable consequence of (*). There are others, too, including:

- we cannot know which fit will help more, or by how much, without doing something equivalent to trying both.

- we cannot know how much better fitting a combined dependence, straight-line in x and t separately, will do, without trying it -- directly or indirectly.

For a simple illustration of these facts -- in two-way table form -- let us turn back to the example of exhibit 26 of chapter 22. Exhibit 1 restates the data; exhibit 2 shows the residuals from four fits.

Fitting a weight tilt alone -- in addition to a constant -- is quite helpful. The residuals become 1/2 to 1/3 of their former size. This teaches us that fitting a rate tilt must help a lot too. Perhaps not as much, perhaps more.

Fitting a rate tilt alone -- in additon to a constant -- is quite helpful. The residuals become 1/4 to 1/8 of their former size. This teaches us that fitting a weight tilt must help a lot too. Perhaps not as much, perhaps more.

In fact, fitting the rate tilt alone does considerably better than fitting the weight tilt alone. Clearly, as any of us can show by inventing an example, the opposite could have been the case.

Fitting a combination of a rate tilt and a weight tilt is also quite effective. For this particular set of data the combined tilt is only about as good as the better of the two separate fits. (We would be sure that the combined fit was poorly chosen, if it were not at least as good as both separate fits.) In other situations, as the hypothetical

exhibit 1 of chapter 23 -- finger sensitivity

The example of exhibit 26 of chapter 22 repeated

Weight	Rate			
	a	b	c	d
1	x	x	x	134
2	x	x	112	130
3	x	x	122	x
4	x	84	110	x
5	x	85	x	x
6	57	81	x	x
7	49	x	x	x

exhibit 2 of chapter 23 -- finger sensitivity

Residuals from the data of exhibit 1 after various fits

Fitting constant and weight tilt

	a	b	c	d
1	x	x	x	-4.1
2	x	x	-12.2	5.8
3	x	x	11.7	x
4	x	-12.4	13.6	x
5	x	2.5	x	x
6	-11.6	+12.4	x	x
7	-5.7	x	x	x

Fitting constant and rate tilt

	a	b	c	d
1	x	x	x	-2.9
2	x	x	2.1	-6.9
3	x	x	12.1	x
4	x	1.1	0.1	x
5	x	2.1	x	x
6	1.1	-1.9	x	x
7	-6.9	x	x	x

Fitting constant only

	a	b	c	d
1	x	x	x	37.6
2	x	x	15.6	33.6
3	x	x	25.6	x
4	x	-12.4	13.6	x
5	x	-11.4	x	x
6	-39.4	-15.4	x	x
7	-47.4	x	x	x

Fitting constant and both tilts

	a	b	c	d
1	x	x	x	-4.1
2	x	x	-1.0	-5.4
3	x	x	11.7	x
4	x	-1.2	2.4	x
5	x	2.5	x	x
6	-0.4	1.2	x	x
7	-5.7	x	x	x

Ranges and midspreads

No tilt	Rate tilt
85.0 and 41.0	19.0 and 5.0
Weight tilt	**Both tilts**
26.0 and 23.8	17.4 and 6.5

example of exhibit 3 shows, the combined fit can -- and sometimes -- be much better than either fit alone.

Finding that fitting a tilt on x helps a lot does not, of itself, tell us that this is a good way to describe the data. There may be either a t such that a tilt on t alone does much better or a t such that a combined tilt on x AND t will do much better. In the first case we can do better with a zero coefficient for x -- in the second the coefficient of x is likely to change -- who knows by how much. We have to remain alert, whenever we find a useful fit, for the possibility that adding another carrier will make a different coefficient better.

23B. <u>A factor can appear repeatedly</u>

When our data starts in terms of two or more numerical variables

$$y, x, t, \ldots$$

we can approach the data in two ways:

- treating all the variables on the same footing, as we did in chapter 13.
- treating one variable as a response, and asking about its relation to the others, as we have done in many chapters.

In taking the latter approach we need to be careful about our choice, both as to its sensibleness and its appropriateness. Way back in chapter 10, for example, where we considered dollar value of U.S. imports and year, it would have shocked most of us to take year as the response to the dollar volume

exhibit 3 of chapter 23 -- handmade example

Hypothetical example where combined fit is very much better than either separate fit

A) The DATA

Weight	Rate 1	3	5	7
1	x	x	x	40
2	x	x	17	35
3	x	x	13	x
4	x	-9	9	x
5	x	-13	x	x
6	-35	-17	x	x
7	-38	x	x	x

Midspread of data: 34
Range of data: 78

B) FIT and RESIDUALS -- rate tilt only

Fit

-37.2	-12.4	12.4	37.2

Residuals

x	x	x	2.8
x	x	4.4	-2.2
x	x	0.6	x
x	3.4	-3.4	x
x	-0.6	x	x
2.2	-4.4	x	x
-0.8	x	x	x

Midspread of residuals: 5.0
Range of residuals: 8.8

C) FIT and RESIDUALS -- weight tilt only

Fit	Residuals			
39	x	x	x	+1
26	x	x	-9	+9
13	x	x	0	x
0	x	-9	9	x
-13	x	0	x	x
-26	-9	9	x	x
-39	+1	x	x	x

Midspread of residuals: 18
Range of residuals: 18

(Cont'd)

exhibit 3 (Cont'd)

D) FIT and RESIDUALS -- both tilts

	Residuals				Fit
	x	x	x	+1	12
	x	x	0	0	8
	x	x	0	x	4
	x	0	0	x	0
	x	0	x	x	-4
	0	0	x	x	-8
	+1	x	x	x	-12
Fit	-27	-9	+9	27	

Midspread of residuals: 0
Range of residuals: 1

of U. S. imports. And in section 11E, where we looked at marriages and motor vehicle deaths and asked how we would fill in a missing value, it was important which variable we took as the response -- important in terms of how our problem appeared to us, certainly not in terms of which "caused" the other. (No one of us believes that either "causes" the other in any direct sense.)

Having got this far, we then need to choose a useful and effective way to express the response. What is the corresponding thing to do about the factor or factors?

- sometimes, when we are just comparing, we only use the factor as a way of cutting the data up. This might be the effective way to use jersey numbers on football players, since the first figure usually tells us about playing position.

- sometimes we want to have a neat, confined fit and want to consider a straight-line dependence on a chosen carrier. Now we have to choose some one expression of the factor as the carrier. Once we have done this, we fit a tilt.

- sometimes we want -- or are forced to use -- a somewhat less constrained fit. Our fitting of tilt and bend together in chapter 14 is an example where we chose two carriers, both functions of x, and sought a combined straight-line dependence.

The situation where we choose two or more carriers determined by the same factor is not at all uncommon. Indeed,

it is important. The process of choosing the two or more carriers is essentially a choice of how we are to express x, a choice with more built-in flexibility than when we choose a single way of expressing y.

* two carriers can be near surrogates *

Let us turn back to the example of fitting both tilt and bend with

$$y = \text{value of U. S. imports}$$

and

$$x = \text{date Anno Domini}$$

which appeared first in section 10C and later at the end of section 22D. As exhibit 18 of chapter 22 makes clear, we can expect a reasonably good fit if we work with the possible fits in the form

$$a_3 + b_3 \cdot x + c \cdot ((x-1923.5)^2 - 0.25)$$

Neatness of appearance, and the practices of high-school algebra would suggest writing the possible fits in the equivalent form (the a and b will be different)

$$a_2 + b_2 x + c\, x^2$$

Would this be a good idea?

Worse choices than this can be made, but only in rather unlikely ways. Exhibit 4 shows how well we can match $x - 1923.5$ with

$$(x^2 - (1923.5)^2)/3847$$

Clearly, if only one of these were to be used, it would take quite precise data to tell which of these was a better fit than the other. Not surprisingly, we will have arithmetic difficulties if we try to fit the two simultaneously.

If this much confusion can occur when a quadratic is put in poorly chosen terms, how much more opportunity for difficulty can we expect when, perhaps because our data is not satisfied with a quadratic, we try the first three powers of x -- or the first four or the first five?

Those who answer "more" will be quite correct. The fitting of polynomials can be easy, if we are careful, but it can also be made as difficult as we may wish. Polynomials

look simple

Sometimes they even are simple, but rarely by accident.

It would be quite wrong to leap to the conclusion that it is unwise to make fits involving two or more expressions of the same factor. This is very far from the truth.

We need to beware of danger, but we can have danger without repeated appearance of a variable. How much danger we have will depend in part -- as we have just seen in one example -- on just how we write down what we are fitting.

exhibit 4 of chapter 23 -- an example

The confusibility of adjusted multiples of x and x^2 when x varies by only a small fraction of itself

Date = x	x - 1923.5	$(x^2-1923.5^2)/3847$
1918	-5.500	-5.492
1919	-4.500	-4.495
1920	-3.500	-3.497
1921	-2.500	-2.498
1922	-1.500	-1.499
1923	-0.500	-0.500
1924	.500	.500
1925	1.500	1.501
1926	2.500	2.502
1927	3.500	3.503
1928	4.500	4.505
1929	5.500	5.508

It is high time we thought harder about the effect of writing down fits in different ways.

* a constant can be a surrogate *

The problem of one variable acting as surrogate for another can come up when one of them is a constant.

Consider the affairs of the (imaginary) DEF Corporation, which has been in existence -- steadily growing -- for 50 years. In the fiftieth year of its existence a careful day-to-day study of its net worth for each of the first 10 days of June gave the results in exhibit 5. Panel B of that exhibit shows two alternative fits:

- the first fit is a constant, which we can usefully regard as

$$a_2 \cdot 1$$

- the second fit is proportional to the age of the company, say

$$b_3 \cdot (\text{age})$$

It is clear that both fits make the residuals much much smaller without making them notably less spread out. Either fit is good; it is hard to choose between them; 1 and age are

quite close surrogates

of each other

for this data

exhibit 5 of chapter 23 -- handmade

Day-by-day net worth of an imaginary corporation

A) THE DATA

Day		Net worth at close of business	
June	1	$50,432,234.21	(Sat)
	2	50,321,458.34	(Sun)
	3	50,510,321.43	(Mon)
	4	50,395,642.28	(Tues)
	5	50,515,243.35	(Wed)
	6	50,520,218.73	(Thur)
	7	50,516,343.27	(Fri)
	8	50,442,527.81	(Sat)
	9	50,347,793.22	(Sun)
	10	50,514,322.43	(Mon)

B) TWO COMPETING FITS

Day		Fit 1	Fit 2
June	1	50,488,390	50,476,060
	2	50,488,390	50,478,800
	3	50,488,390	50,481,540
	4	50,488,390	50,484,280
	5	50,488,390	50,487,020
	6	50,488,390	50,489,760
	7	50,488,390	50,492,500
	8	50,488,390	50,495,240
	9	50,488,390	50,497,980
	10	50,488,390	50,500,720

(If we had net worths over 10 years instead of 10 days, l and age would not be close surrogates.)

23C. Posses

Before we start looking at different ways of writing down the same fit, we ought to look -- and learn to look effectively -- at what possibilities our process of fitting has considered, not just at the one representative that it selected as OUR fit to THIS data.

As we do this, we need a word for such objects, such collections of possible fits. Different authors -- and speakers -- have different practices in how they use words. To some "model" or "specification" would be reasonable -- more often merely almost reasonable -- words to use for such things, but others would disagree (with violence). Invention seems better than compromise, so we shall introduce a new word, calling a collection of possible fits a

posse

(and thus satisfying many punsters in two ways).

Turning back to chapter 22, where we let the numbers fade away, and where the versions of both x and t were often taken as equally spaced, we were concerned with a sequence of posses, beginning with

$$A + f(x) + g(t)$$

where A was any constant, f(x) was any function defined for the four values taken on in the data by x, and g(t) was any function defined for the seven values taken on in the data by t.

The second posse was

$$A + f(x) + c_1 t$$

where A and f(x) were as before, and c_1 was any constant.

The third posse was

$$A + b_1 x + c_1 t$$

where A, b_1, and c_1 are any constants.

We know how to make a picture of any one of the fits that make up these posses.

Any two-way plot made up of two sets of parallel straight lines -- each line labelled with a version of that factor -- represents some fit belonging to the first posse. To talk about this posse is to talk about all two-way plots of this kind. Not one, not three, not a thousand, but ALL.

Consider any two-way plot made up of two sets of parallel straight lines, for which

- each line is labelled with a version of the corresponding factor.
- the lines corresponding to factor t are equally spaced -- and labelled with equally spaced numbers, in order from one extreme version to the other.

The second posse consists of ALL fits that can be represented by such two-way plots. Every fit that belongs to the second posse clearly belongs to the first posse. The second posse is thus a part of the first posse.

The third posse consists of ALL fits that can be two-way plotted by two sets of parallel straight lines for which

- each line is labelled with a version of the corresponding factor.

- the lines corresponding to each factor are uniformly spaced -- the two families can have different spacings -- and each family is labelled with equally spaced numbers in order from one extreme version to the other.

Clearly the third posse is part of the second one.

23D. Writing the same thing down two ways

Let us now look at the third of these posses -- all fits of the form

(*) $$A + b_1x + c_1t$$

where A, b_1, and c_1 are any constants. We have already noticed that for fixed A, b_1, and c_1, this fit is exactly equivalent to

$$A + (b_1-c_1)x + c_1(t+x)$$

as is easy to check by multiplying out and cancelling.

We now notice that this last form shows our fit to be an example of the posse

(**) $$A + b_2x + c_2(t+x)$$

with

$$b_2 = b_1 - c_1$$

and

$$c_2 = c_1$$

This means that posse (*) is part of posse (**).

Now look at any fit belonging to posse (**). It takes the form

$$A + b_2 x + c_2(t+x)$$

which is exactly equivalent to

$$A + (b_2+c_2)x + c_2 t$$

clearly an example belonging to posse (*). Hence posse (**) is part of posse (*). Hence they consist of exactly the same fits.

There are three sorts of things we need to know in this situation:

- that the two posses are the same, so that we set ourselves the same fitting problem by selecting either as what we will consider -- as the collection of fits among which we are to choose, guided by the data.
- that any fit in one posse is exactly equivalent to a fit in the other, so we can work out our fit in either posse, transferring it to the other one if that is where we want to have it.

- how the details of the fit in one posse are related to the details of the fit in the other posse. We need this both to be able to change back and forth in practice and to be able to understand better what the details mean.

* what do the details mean? *

We started on the third of these points in the last chapter, when we noticed that

$$A + b_1 \cdot x + c_1 \cdot t$$

was exactly equivalent to

$$A + (b_1 - c_1) \cdot x + c_1 \cdot (t+x)$$

We did not stress what happened among the coefficients b_1, c_1, $b_1 - c_1$ and c_1 then. It is time to stress just this now.

Let us simplify appearances without changing essential by assuming that $A = 0$. Then our exactly equivalent fits are

$$b_1 \cdot x + c_1 \cdot t$$

and

$$(b_1 - c_1) \cdot x + c_1 \cdot (t+x)$$

Passing from the former to the latter, instead of writing our fit in terms of x and t, we wrote it in terms of x and $t+x$.

We kept x , and changed t . This did NOT keep the coefficient of x the same and change that of the other carrier. Quite the reverse! The coefficient of $t+x$ stayed the same as that of t , while that of x changed.

This may appear backward at first, but repeated trials of the algebra -- especially with numbers for coefficients -- will convince most of us that this is the way things must be. One conclusion is now inescapable. The coefficient of x cannot be thought of as telling us

about how y appears to be related to x

If it told us that, it could not matter whether t or $t+x$ was the other carrier. Instead, we have to think of this coefficient as telling us

about how y appears to be related to x in the presence of certain other carriers

Let us suppose that "response" has its strongest meaning, so that we can properly think of changing x or t or both and watching the effect on y . If

$$y \text{ is fitted by } b_1 \cdot x + c_1 \cdot t$$

it is natural -- but wrong -- to think of b_1 as telling us about

how y responds to x

Were this right, the equivalent statement

$$y \text{ is fitted by } (b_1-c_1)\cdot x + c_1\cdot(t+x)$$

would say that b_1-c_1 tells us about just the same thing as b_1 does, whatever c_1, something which just cannot be.

The only escape is to recognize that in

$$y \text{ is fitted by } b_1\cdot x + c_1\cdot t$$

b_1 has to be thought of as telling us -- probably incompletely -- about

how y responds to x when t is fixed

and that, in

$$y \text{ is fitted by } (b_1-c_1)\cdot x + c_1\cdot(t+x)$$

b_1-c_1 has to be thought of telling us -- probably incompletely -- about

how y responds to x when $t+x$ is fixed

Now our two different numbers -- b_1 and b_1-c_1 -- are thought of as telling us about two different things. This is as it should be.

* "held fixed", you say? *

That b_1 and b_1-c_1 have different meanings is certain. That "while ____ is held fixed" is the right way to describe the difference is very dubious.

There can be many reasons why we cannot hold the other carrier -- or carriers -- fixed. The simplest example arises when t is x^2 or $(x-1923.5)^2$. Clearly, holding either x^2 or x fixed while changing the other is not a relevant idea. This always happens for repeated appearances of a variable, but can also easily happen when the connection between the carriers occurs only in our set of data.

Exhibit 6 shows a hand-made example. We can never hold either x or t fixed

in the data

while moving the other, yet the joint fit is the best description of the response we can give.

It is much better to think -- and sometimes even to say:

b_1 tells us how y responds to x after we have adjusted -- in a straight-line way -- for any simultaneous change in t

In using this somewhat more complicated -- and much more realistic -- form of statement and thought, we also need to be careful about the nature of the straight-line adjustment. It is not the adjustment that we would make if we were to fit t alone. The presence of x has affected the coefficient of t, just as the presence of t has affected the coefficient of x. The adjustment we make is

that adjustment which can be best combined with a multiple of x to give a close fit

exhibit 6 of chapter 23 -- handmade

A hand-made example where "When_____is fixed" makes no sense

x \ t	1	4	7	13	27	63
1	.	5	.	.	.	.
2	3	.	.	.	.	.
3	.	.	.	.	30	.
4	.	.	.	.	.	67
5	.	.	12	.	.	.
10	.	.	.	23	.	.

Note: x+t is a perfect fit.

* the third caution *

In the last few paragraphs we inserted "-- possibly incompletely --" in some very crucial statements. Just a little earlier we had said "appears to be related". Clearly we are being cautious. Equally clearly we would like not to have to be cautious.

We must ask whether we need to be cautious.

Exhibit 7 gives a simple hand-made example. There is no doubt that

$$10x + 10t - 15$$

does a useful job of fitting y. It reduces a midspread of 30 (and range of 80) to a midspread of 10 (and range of 40). Indeed, "10x" does tell us -- in an average sense -- how y responds to a change in x .

Only in an average sense, however, since if t is held fixed at 1, changing x has no effect on y , while if x is held fixed at 5, y changes by 20 units -- not 10 -- for each unit change in x .

* the three shadows *

In dealing with even the simplest fits involving two (or more) carriers -- even when the fit does very well in shrinking the residuals -- there are at least three cautions that we need to keep in mind:

- either or both of the carriers may be effective because it is --- at least in the data before us ---

exhibit 7 of chapter 23 -- handmade

Another example

A) The DATA

t	x 1	2	3	4	5
1	25	25	25	25	25
2	25	30	35	40	45
3	25	35	45	55	65
4	25	40	55	70	85
5	25	45	65	85	105

B) An ELEMENTARY ANALYSIS

t	x 1	2	3	4	5	eff	fit	(str) (fit)
1	20	10	0	-10	-20	-20	25	(25)
2	10	5	0	-5	-10	-10	35	(35)
3	0	0	0	0	0	0	45	(45)
4	-10	-5	0	5	10	10	55	(55)
5	-20	-10	0	10	20	20	65	(65)
eff	-20	-10	0	10	20	45	00	
fit	25	35	45	55	65	00	-45	
(str-fit)	(25)	(35)	(45)	(55)	(65)			

C) FIT by COMBINED STRAIGHT-LINE DEPENDENCE

$$10x + 10t - 15$$

D) REMARK

If we calculated comparison values and made an extended fit, we would get a perfect fit.

closely related to some variable we would have been better off to have used (if we could). **Surrogates.**

- the coefficient of a carrier can only be interpreted in terms of what other carriers are in the fit -- and are, in the simplest cases, to be held fixed when thinking about that coefficient. **Co-carriers.**

- the fit, though useful, may be far from perfect; its descriptions of behavior may only apply on the average -- and not to particular cases of holding the other carriers constant. **Incompleteness.**

For all that any fit is plagued by these three shadows, fitting -- as we have seen -- is generally useful and often essential.

In particular, it is only by fitting that we get hold of residuals, and it is only by residuals that we can attack incompleteness.

Equally, it is only by alternative fittings that we learn how effective a surrogate one of our carriers would be for some possible carrier we did not use.

The question of co-carriers, on the other hand, is clearly posed once we see the details of the fit. We can easily see what happens if we change the co-carriers while preserving the posse.

After we have done some of this, however, we may have to think hard about which set (or which set of sets) of co-carriers corresponds best to the question we may want to ask.

Similarly, once we have found out what other possible carriers might be surrogates for by one of the carriers we use, we may have to think hard about which one we ought to use. (For an example, go back to chapter 16 and the Arizona temperatures, where we might have to decide whether north-south or higher-lower was the important carrier in comparing Yuma, Phoenix, Prescott, and Flagstaff.)

We can deal with these three shadows -- more or less completely according to our problem. To do this, we need BOTH to look at the numbers -- both at the residuals of our fit and at other possible fits -- AND to think hard about which choices of carriers match both what we want to know and the mechanisms of the real world.

23E. <u>But is this exploratory?</u>

The discussion of this chapter may have seemed a little out of keeping with what has gone before. After all, our main thrust throughout has been "clearing the trail and looking deeper". So long as we peeled off the upper levels of phenomena well enough to reveal what lay underneath, we appear to have given little attention to what was peeled away. Rather we have been eager to see what new was revealed.

This is all true, but why have we acted in this way? Because our concern was with quantitative detective work -- and our focus was on those aspects of it that are likely to give the practitioner trouble.

Throughout the greater part of all we have done, the

appearance of what was peeled away

has been clear. We may not have understood why certain Nebraska counties, adjusted for overall behavior, showed time trends (in the % Democratic for President), but we have had no difficulty in saying what such a time trend was. Interpretation may have been uncertain, but appearance seemed clear. (This is the pattern of many of a conventional detective's clues.)

There was an earlier exception to this pattern. When we came to find tilts, we were careful to make clear that a tilt need not be a slope, that in thinking about an apparent tilt we must be careful not to believe that it must also be an apparent slope. We made that emphasis for the same reason that we are emphasizing rather similar cautions here: Both are situations where the quantitative detective can too easily go astray by putting false values on indications.

The "Quantitative Eye" can go wrong by failing to look well after the road has been cleared, or by failing to do a good enough job in clearing the road, or by failing to recognize what it is that has been cleared from the road. We need to try to help him to avoid each of these failings, distributing our efforts to match his needs.

What we have been doing in this chapter is, then, just another part of our main mission.

23F. Unplanned data

Much exploratory data analysis is analysis of unplanned data. Not all matters of interest are subject to carefully planned comparative experiment. Astronomy, many aspects of biological evolution, international trade, the sales of a new brand of motor car, and the labor relations of a major corporation with uniform, nationwide operations are all essentially immune from experiment.

Many difficulties which arise in the analysis of unplanned data are inescapable. Many economic series tend to grow at about the same rate per year. As a consequence, measuring the rate of growth of some one series that interests us tells us little about which -- if any -- of the many more basic series with about the rate of growth measured for our series is an important driving force behind the series that interests us. No change in techniques of analysis is likely to do anything appreciable to ease this sort of difficulty. More drastic actions are likely to be essential, perhaps:

- bringing in some quite different sort of data, to be used in combination with the data before us.
- establishing, by some believable route, that certain simplifying assumptions are not only helpful but reasonable. (The latter is usually much harder to establish.)
- admitting that one is not going to find out -- at least for the present.

To say that such action may be necessary is no ground for not striving to analyze what data we have. The conventional detective does not hesitate to search a room because he may not find a clue -- nor does he hesitate because many detectives have searched many rooms without finding a clue. It would be just as wrong, of course, for a quantitative detective to insist in advance that his analysis will certainly answer the question as for the conventional detective to insist, equally in advance, that his search of the room will certainly find a clue.

Unplanned data raises many difficulties -- some with which we are familiar, some we are just beginning to recognize, at least a few that we have not yet come to recognize in even the slightest degree -- but that is the way of the real world. Those who would explore geography must be prepared to face the jungle -- or the desert -- or the ice cap. Those who would explore data must be equally ready to face their own difficulties.

Any serious reader would like a painlessly memorizable package listing all possible difficulties and giving one or two sure cures for each. How else could a reasonable man dream?

There are various reasons why there will be no such package here. Some of these are:

- we do not know all difficulties.
- we do not know sure cures for many known difficulties.
- when we do know something that usually cures a given difficulty, it is not likely to be easily described or memorized.
- this is an introduction.

At this point, all we can do is to combine heartfelt but general warnings with equally heartfelt shouts of "c'mon in, the water's fine!".

*****

23H. How far have we come?

In this chapter we have tried to be realistic about the most important ways in which fitting can be confusing. The two essentials of a balanced view are simple:

- yes, fitting can be confusing, and we must be careful.

- yes, there is nothing like fitting for getting ahead.

While it would be nice -- but so, so, surprising -- if fitting would solve our problems for us almost automatically, we ought to be grateful for the many useful things it does do for us. The efforts required to learn about -- and constantly allow for -- its difficulties and confusions are efforts very well spent.

In looking at fitting simple expressions involving two carriers -- involving straight-line dependence on each of two quantities that describe the situation in which the response occurs -- we can learn quite a lot by thinking of our data as a more-or-less complete two-way table whose factors are specified by the carriers.

Looking back to the examples of the last chapter, we find simple examples where one carrier can act as a more or less complete surrogate for another. The possibility of this needs always to be before our eyes.

We have earlier seen examples -- like the fitting of tilt and bend -- where two or more carriers express the same

variable. There is nothing wrong with this; it is a standard situation, in which regarding the data as a sparse two-way table may help. If either of two carriers can serve as a surrogate of the other -- as happens too often to incautious analysts -- we must again be careful. Better choices of equivalent carriers is likely to get us out of trouble here.

Once we start talking of equivalent carriers, we are really talking about families of possible fits, not about just the one fit we chose. Having both a name for -- and some facility in thinking about -- such posses is important.

There are equivalent ways to write down both simple fits and simple posses. When we compare one with another, especially when one carrier is the same and the other different, we find their relationship is almost backward to what we naively expect. The coefficient of the unchanged carrier changes, while the coefficient of the changed carrier can be the same.

This means that we must interpret the coefficient of any one carrier as being "in the presence of all the other carriers" -- as telling about the effect of this carrier "with all the other carriers held fixed". Again continuing caution is essential.

Finally, we must remember that fits are, by their nature, incomplete descriptions. They may -- if we are wise about selecting the response and the factors, lucky about surrogates, and careful about co-carriers -- tell us

on the average

what the effect of one carrier is on the response when the other carriers are held fixed. This need not be the same as the effect of that carrier when the other carriers are held fixed at some particularly interesting values. To examine how incomplete our fit is, and how it ought to be improved, we need to make the fit and look at the residuals.

Surrogacy, co-carriers, incompleteness -- we must always expect our fits to be plagued by these three diseases. We need to be alert to detect -- and do what we can to remedy -- the problems that they pose in each special situation. For all these diseases, there is no substitute for fitting, especially when dealing with unplanned data.